Title

Predicting Titanic Survival Rates: A Comparison of AdaBoost, XGBoost, and Random Forest

Author

WU, TONGCHANGYU

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Predicting Titanic Survival Rates:

A Comparison of AdaBoost, XGBoost, and Random Forest

A dissertation submitted in partial satisfaction of the

requirements for the degree Master of Applied

Statistics & Data Science

by

Wu TongChangYu

2024

ABSTRACT OF THE DISSERTATION

Predicting Titanic Survival Rates:

A Comparison of AdaBoost, XGBoost, and Random Forest

by

Wu TongChangYu

Master of Applied Statistics & Data Science

University of California, Los Angeles, 2024

Professor Ying Nian Wu, Chair

The factors influencing survival rates during disasters had always been an important subject of research. With the rise of machine learning, predictive modeling has improved significantly. This paper presented a comparative analysis of three Machine Learning models—XGBoost, Random Forest, and AdaBoost—trained using well-established libraries to predict the survival probabilities of passengers on the Titanic. We used a well-known dataset from the Titanic disaster, containing passenger information and whether they survived. After data preprocessing and model tuning, Random Forest showed the highest accuracy, suggesting its potential for improving survival predictions in disaster rescue operations.

The dissertation of Wu TongChangYu is approved.

Nicolas Christou

Oscar Madrid Padilla

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The sinking of the RMS Titanic on April 15, 1912, stood as one of the most tragic and well-known maritime disasters in history. This disaster resulted in the deaths of over 1,500 passengers and crew members, becoming a subject of extensive study and analysis. The Titanic symbolized the engineering marvels of the early 20th century, and with its luxury and advanced design led it to be considered the "unsinkable ship." However, the tragic collision with an iceberg during its maiden voyage exposed significant flawed in the maritime safety practices and design assumptions of the time.

On April 10, 1912, the Titanic departed from Southampton, England, with New York City as its ultimate destination. Measuring approximately 269 meters in length, 28 meters in width, and boasting a gross tonnage of 46,328 tons, it was the largest vessel at that time. It offered luxurious amenities for its first-class passengers, including a grand staircase, swimming pool, library, fine dining rooms, and lavish cabins, while also providing relatively simple but still comfortable accommodations for second and third-class passengers. On board were 2,224 people, including wealthy individuals, ordinary passengers, and many immigrants seeking new opportunities. This diversity of passengers and their social classes provided a wealth of data for analyzing the human factors involved in the disaster.

At 11:40 PM on April 14, lookouts spotted an iceberg and attempted to avoid it, but the starboard side of the ship scraped against the iceberg, causing significant damage to the underwater hull, the watertight compartments were breached, leading to rapid flooding of

the ship's interior. The subsequent evacuation process revealed numerous inadequacies in maritime safety protocols of the time. There were not enough lifeboats to accommodate all passengers and crew, and the evacuation was chaotic and inefficient. Many lifeboats were launched only partially filled, and the prioritization of passengers varied, although women and children were generally given priority. The nearby SS Californian failed to respond to distress signals, while the more distant RMS Carpathia rushed to the scene, arriving several hours later to rescue 705 survivors.

This incident had highlighted the importance of quickly and accurately predicting survivor characteristics for effective rescue planning during major disasters. In recent years, detailed data on passengers and crew had enabled researchers to use advanced analytical techniques to study the event in depth. Traditional statistical analysis methods usually relied on manually constructed mathematical models, but machine learning algorithms could automatically extract features from data and build models, reducing human bias and improving prediction accuracy. With the explosion of data and the enhancement of computing power, machine learning had shown its superior modeling and prediction capabilities in various fields, leading to widespread application. Therefore, an effective approach to studying the Titanic dataset was to use machine learning to predict survival outcomes based on different passenger attributes. By training on historical data, researchers could uncover patterns and better identify key factors influencing survival.

In this study, we compared three different machine learning models: XGBoost, Random Forest, and AdaBoost, to determine which one is most suitable for prediction. Machine learning algorithms could thoroughly analyze patterns within the Titanic dataset, including

variables such as age, gender, cabin class, ticket fare, cabin location, and family relationships. This research not only enhanced our historical understanding of the Titanic disaster but also demonstrated the remarkable ability of machine learning to extract meaningful patterns from complex, large-scale datasets. By using these advanced techniques, we could better understand the critical factors affecting survival in maritime disasters and provide data-driven support and insights for risk management in similar future situations. Combining historical analysis with modern computational methods offered a novel, comprehensive approach to studying this famous maritime tragedy, contributing theoretical value for prevention and preparedness.

This study aimed to improve disaster management and risk assessment capabilities by using machine learning models to simulate and predict disaster scenarios and survivor characteristics, providing data support for developing more effective disaster response strategies. Governments and relevant organizations could conduct this approach to improve disaster management, risk assessment, maritime safety, and emergency response capabilities. Examining key survival factors not only served as a reference for future emergency rescue operations and the formulation of targeted and efficient rescue strategies but also helped improve modern maritime safety measures and emergency response mechanisms. Optimizing evacuation procedures and creating more efficient and fair evacuation plans ensured that passengers could evacuate quickly and safely in emergencies, reducing casualties and providing practical application value.

Predicting key factors that affected survival rates through machine learning models enabled shipping companies and ship designers to better allocate life-saving equipment, ensuring a rational distribution and adequate reserve of lifeboats and other safety

equipment on board, enabling timely and effective assistance during disasters. Moreover, this research promoted interdisciplinary research and application. The Titanic dataset provided a classic case study for machine learning and statistical analysis, offering empirical data support for social science and behavioral studies, and helping to understand human behavior and decision-making in crisis. Furthermore, integrating historical events with engineering design offered insights for improving the safety design and management processes of modern engineering projects, providing a unique perspective for both historical and engineering research.

The remainder of the paper was structured as follows: Section 2 provided an overview of the methodology employed in this study. Following that, Section 3 delved into the details of the dataset, including all its features and how we processed it. Moving on to Section 4, we conducted an exploratory data analysis (EDA) to gain deeper insights into the dataset. Section 5 discussed different models that were compared in this study, elaborating on their predictive capabilities after parameter tuning. Finally, Section 6 presented the conclusions drawn from our analysis, along with suggestions for future research directions. This was followed by a list of references and an appendix containing supplementary information.

CHAPTER 2

Methodology

In this section, we will delve into these three supervised Machine Learning (ML) algorithms we used to predict survival rates: AdaBoost, Random Forest, and XGBoost. Each algorithm its owns unique strengths, contributing to the development of robust and accurate predictive models.

By comparing the unique capabilities of AdaBoost, Random Forest, and XGBoost, we could determine which algorithm offered the most accurate and reliable predictions for survival rates. Each of these models brought distinct advantages to the table, enabling us to evaluate their performance comprehensively and chose the one that best met our requirements.

2.1 AdaBoost (Adaptive Boosting)

Boosting, also known as enhanced learning, is an important ensemble learning technique that can transform weak learners with only barely better than random prediction accuracy into strong learners with high accuracy. This method offers a new and effective way to design learning algorithms, especially when it is challenging to construct strong learners directly. AdaBoost's strength lies in its adaptive approach: the weights of incorrectly classified samples by the previous classifier are increased, while the weights of correctly classified samples are decreased. These adjusted sample weights are then used to train the next weak classifier. In each iteration, a new weak classifier is added until a

predetermined low error rate or a maximum number of iterations is reached, ultimately constructing a strong classifier.

Specifically, AdaBoost's fundamental principle is iteratively training multiple weak classifiers to build a strong classifier. Initially, all training samples have equal weights. The first weak classifier is trained, and then sample weights are adjusted based on its error rate, increasing the weights of misclassified samples to give them more attention in the next round of training. Each iteration involves training a new weak classifier and assigning it a weight based on its error rate, with better-performing classifiers getting higher weights. This process repeats until the number of iterations or the classifier's performance reaches the desired level. Finally, all weak classifiers are combined with their respective weights, forming a stronger classifier with better overall performance. Therefore, AdaBoost significantly improves classification accuracy, particularly for difficult-to-classify samples.

AdaBoost offers several advantages that make it a powerful tool for classification tasks. Firstly, it significantly enhances classification accuracy, especially for difficult-to-classify samples, which is achieved by adaptively adjusting the weights of the samples, ensuring that misclassified samples receive more attention in subsequent training iterations. Secondly, it is easy to use and could be paired with various weak classifiers with minimal need for parameter adjustment. It adapts the assumed error rate based on feedback from the weak classifiers, leading to efficient performance. Furthermore, it can work with simple weak classifiers without the need for extensive feature selection, and it effectively avoids overfitting.

2.2 Random Forest

Random Forest is an ensemble algorithm consisting of many decision trees operating in parallel, a method known as Bagging. Each decision tree in the forest operates independently and without correlation to the others. When handling classification tasks, each tree in the forest independently evaluates and classifies a new input sample. The final result would be obtained by voting or taking the mean to ensure that the model achieves high accuracy, generalizes well to new data, and maintains good stability.

Fig 1 An example of Random Forest

The strength of Random Forest lies in its characteristics of its use of "randomness" and its "forest" structure. The former helps prevent overfitting, while the latter significantly enhances prediction accuracy. Regarding randomness, on the one hand, through sample

perturbation, specifically utilizing Bootstrap Sampling, which contributes directly to introducing diversity into the dataset by allowing the samples from the original training set to be included multiple times in the sampling set. On the other hand, For each node in a decision tree, a random subset of features is selected, and the best feature from this subset is used for splitting. This repeated random selection increases the variability among the trees. As for forest, diverse decision trees are obtained by training multiple sampling sets, and then conducting a vote or taking the mean, resulting in higher prediction accuracy than most single algorithms.

Random Forest model offers numerous advantages. Firstly, since each decision tree can be independently generated, supporting parallel computing and contributing to fast training speeds. The use of bootstrapping and random feature selection helps control overfitting. Furthermore outstanding data adaptability is also exhibited by enabling handling both discrete and continuous data, as well as data with nonlinear relationships. Moreover, it is capable of high-dimensional feature datasets without the need for feature selection or data normalization. Also, The use of bootstrapping and random feature selection helps control overfitting. Additionally, it is also effective with imbalanced datasets, balancing classification errors efficiently. Overall, it demonstrates exceptional performance and applicability across a wide range of applications.

2.3 XGBoost (eXtreme Gradient Boosting)

XGBoost stands for “extreme gradient boosting”(Chen & Guestrin, 2016) and is a scalable, powerful gradient-boosted decision tree (GBDT) machine learning library that provides

parallel tree boosting and excels in regression, classification, and ranking problems. The objective function of XGBoost is:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

$$\downarrow \qquad\qquad \downarrow$$

Training loss Complexity of the Trees

$$where\ \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

The objective function of XGBoost has two main components:

1, The first component measures the difference between the predicted scores and the actual scores.

2, The second part is the regularization term, which helps prevent overfitting.

The regularization term includes:

γ: a parameter that controls the number of leaf nodes.

T: the number of leaf nodes.

w: the score of each leaf node.

λ: a parameter that ensures the scores of the leaf nodes are not too large to prevent overfitting.

The fundamental principle of the algorithm is to iteratively add decision trees and perform feature splits, gradually constructing a complete tree. With each tree added, a new function is learned to fit the residuals from the previous prediction. After k iterations, training results is k decision trees. To predict the score for a sample, identify the corresponding leaf nodes based on the sample's features in each tree, where each leaf

node corresponds to a score. The final prediction for the sample is the sum of these scores from all k trees.

XGBoost uses regularization to prevent overfitting and improves generalization performance. It employs several strategies to prevent overfitting, such as: regularization term, shrinkage, and column subsampling. Although the trees are built sequentially, nodes at the same level within a tree can be processed in parallel. Specifically, for each node, the selection of the best split points and the calculation of candidate split point gains are conducted in parallel using multi-threading, which accelerate the training process. Moreover, it is designed to handle sparse datasets and missing values, further optimizing computational performance.

CHAPTER 3

Data

Our dataset consisted of 12 variables and over 1,300 rows, containing detailed passenger information. This included various attributes such as Age, Sex, Pclass, Fare, and more. Table 1 in Section 3.1 provided a comprehensive overview of the dataset, outlining each variable and its respective characteristics. This detailed dataset formed the basis for our analysis and helped in understanding the factors influencing survival rates.

3.1 Feature

Variable	Definition	Notes
PassengerID	Passenger ID	
Name	Name	
Sex	Sex	Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
Survived	Survival	0 = No, 1 = Yes
Age	Age in years	

Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
SibSp	Number of siblings / spouses aboard the Titanic	
Parch	Number of parents / children aboard the Titanic	
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Table 1 Features of Titanic dataset

3.2 Data Preprocessing

Initially, there were no duplicate values in the data, but numerous outliers existed. These outliers typically represented various extreme cases that fell outside the scope of the research and did not contribute to the objectives, so we decided to remove them entirely. Additionally, some variables such as Pclass, Age, Sibsp, and Parch had missing values. To ensure model consistency in the model, we filled these missing values with the median of the respective variables. During the analysis, the Fare variable exhibited severe skewness, which could negatively impact the model due to the imbalanced data.

Therefore, during the exploratory data analysis (EDA) stage, we performed data transformations to make the overall distribution more uniform.

There were also some variables related to basic passenger information, such as Passenger ID, Name, and Ticket. These variables were primarily used for the company's data tracking purposes and essentially useless for the model. Furthermore, Cabin and Pclass and Fare exhibited some redundancy, so removing them simplified the model training process and improved computational efficiency. Finally, before inputting the data into our model, we applied dummy encoding to convert all categorical variables into numerical variables, which was the required input format for our model.

CHAPTER 4

Exploratory Data Analysis (EDA)

4.1 Survival Rates Distribution by Numerical Variables

In the first part of our exploratory data analysis (EDA), we aimed to identify trends between continuous variables and survival rates. This involved examining how various continuous factors, such as SibSp, Parch, Age, and Fare correlate with the likelihood of survival. By analyzing these relationships, we hoped to uncover patterns and insights that might explain differences in survival rates.

4.1.1 Number of Siblings/Spouses and Survival Rates

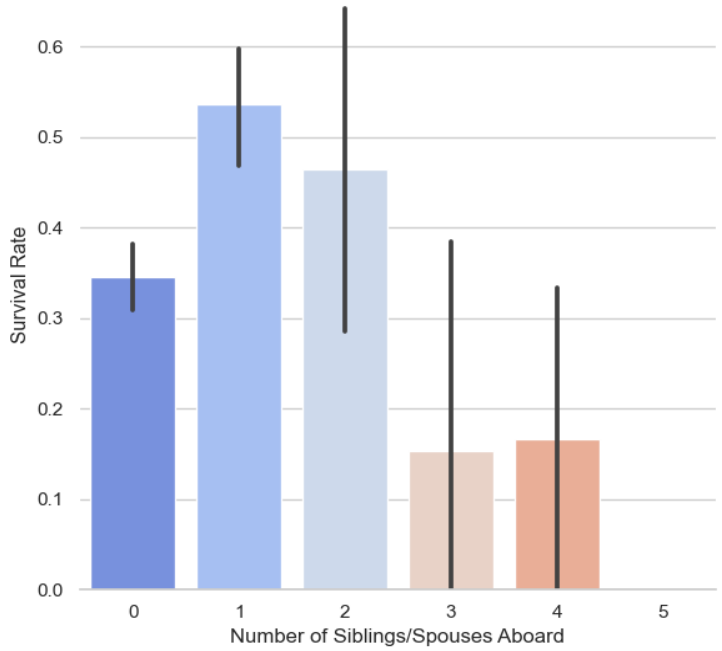

Fig 2 Impact of Siblings/Spouses on Survival Rates

The plot above showed that the number of siblings/spouses (SibSp) a passenger had on board significantly impacted their chance of survival. Passengers traveling alone (SibSp = 0) had a survival rate of 34.5%, while those with one sibling/spouse had the highest survival rate at 53.4%. However, the survival rates dropped to 46.4% for those with two siblings/spouses and fell sharply for those with more than two.

Several factors explained this trend. Passengers traveling alone or those with fewer family members could make quicker and more effective decisions during the evacuation, faced fewer coordination challenges, and had better access to lifeboats. In contrast, larger families had to manage more complex social dynamics and collective decision-making processes, which likely slowed their response and reduced their chances of survival. Additionally, the limited lifeboat capacity and the "women and children first" policy often meant that large families were separated, further decreasing their survival rates. This analysis highlighted the importance of family composition on survival rates and underscored the critical role of social dynamics during emergencies.

4.1.2 Family Size (Parch) and Survival Rates

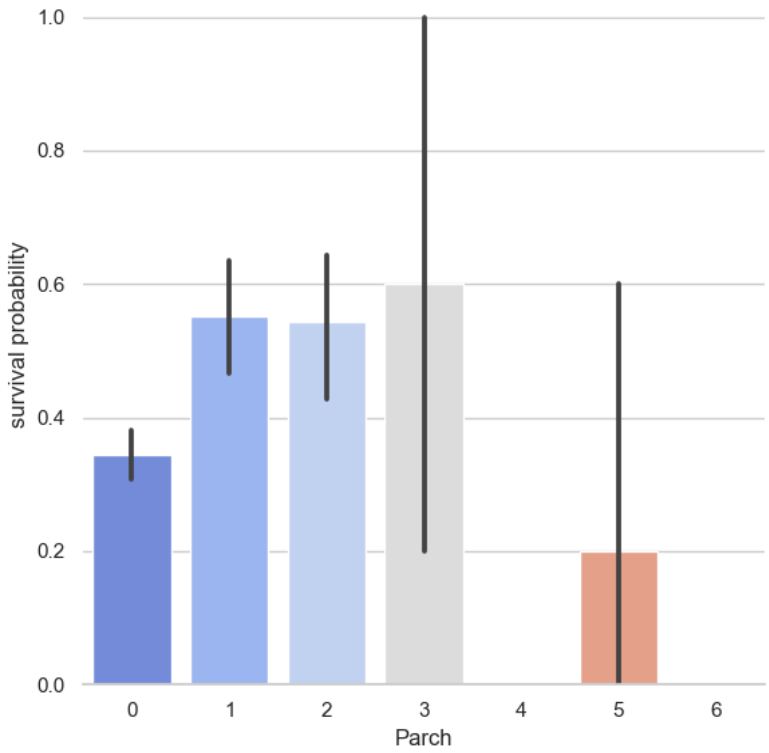

Fig 3 Impact of family size (Parch) on Survival Rates

The figure above illustrated that family size, based on the number of parents/children (Parch), had a significant impact on survival rates. Small families (Parch = 1 or 2) had higher survival chances than single passengers (Parch = 0), medium-sized families (Parch = 3 or 4), and large families (Parch = 5 or 6). This advantage likely came from better support and coordination among small families, helping them secure lifeboat spots more effectively. Single passengers, lacking immediate support, had lower survival chances. Medium-sized families exhibited a lot of variation in survival rates, likely due to differences in family dynamics and decision-making during the crisis. Large families faced significant difficulties because managing more people and the higher risk of separation

made survival more challenging. The high standard deviation in survival rates for passengers with three parents/children highlighted the crucial role of family cohesion and individual actions in such emergencies.

4.1.3 Age Distribution and Survival Rates

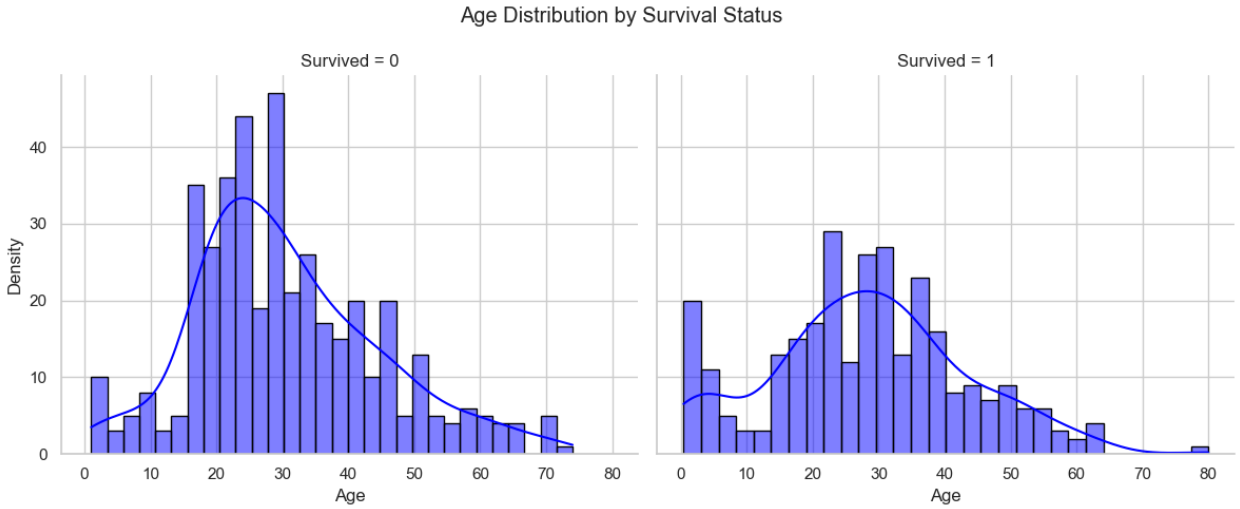

Fig 4 Age distribution by Survival status

The age distribution of Titanic passengers appeared to follow a tailed, possibly Gaussian distribution. There were notable differences between the ages of those who survived and those who did not. Young passengers had a noticeable peak in survival rates, while old passengers, especially those aged 60-80, had lower survival rates. This suggested that although "Age" wasn't directly correlated with "Survival" overall, certain age groups had different chances of survival. Specifically, very young children had a higher chance of survival.

4.1.4 Age Density Distribution of Survival Status

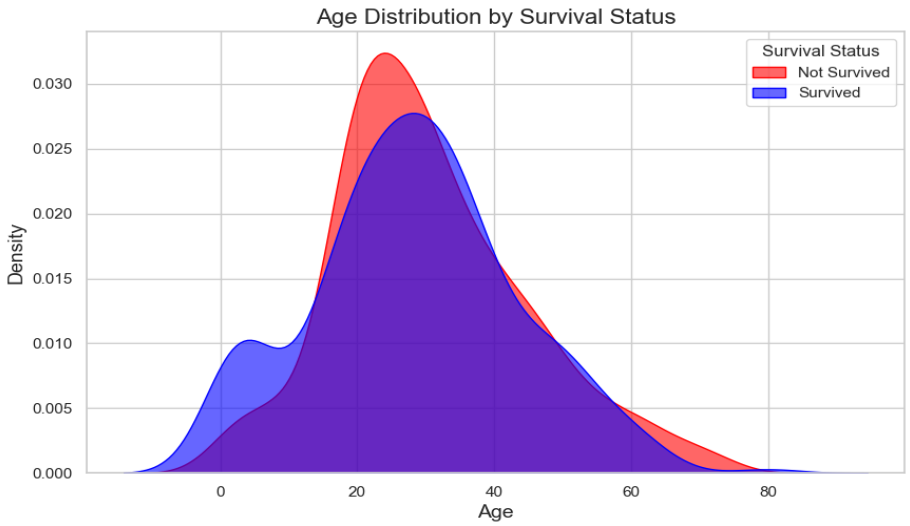

Fig 5 Age density distribution of Survivors vs. non-Survivors

When we overlaid the age density distributions of the survivors and non-survivors, a clear peak emerged for children aged between 0 and 5 years, indicating they had a significantly higher survival rate compared to other age groups. Moreover, the distribution showed that older passengers, particularly those aged 60-80, had much lower survival rates, which showed that while age in general might not predict survival, specific age groups had distinct chances of surviving. Very young children were more likely to survive, whereas older adults faced greater challenges. This pattern underscored the importance of considering age-specific trends when analyzing the factors that influenced Titanic survival rate.

4.1.5 Fare Distribution and Survival Rates

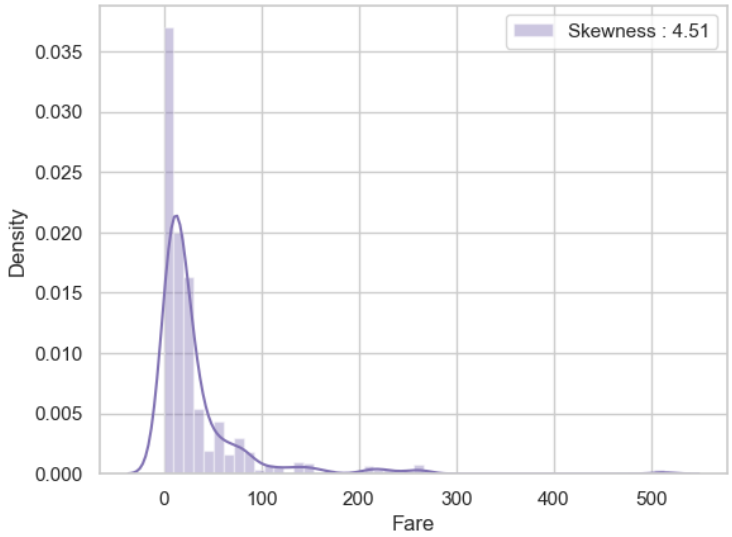

Fig 6 Fare distribution

Looking at the fare distribution, it was highly skewed, indicating an imbalance in the dataset. This imbalance could create issues during the modeling process since some models were sensitive to such skewed data. To address this, we applied a log transformation, which significantly reduced the skewness. The next step was to examine how this transformed fare data correlated with the survival rates.

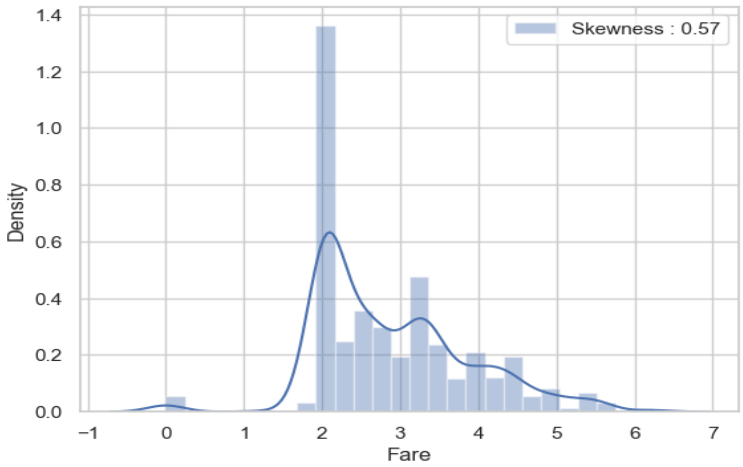

Fig 7 Fare distribution after log transformation

We could clearly observe that the skewness of the data is significantly reduced. Our next step was to see its correlation with our target variable survival rate.

4.1.6 Fare and Survival Rates

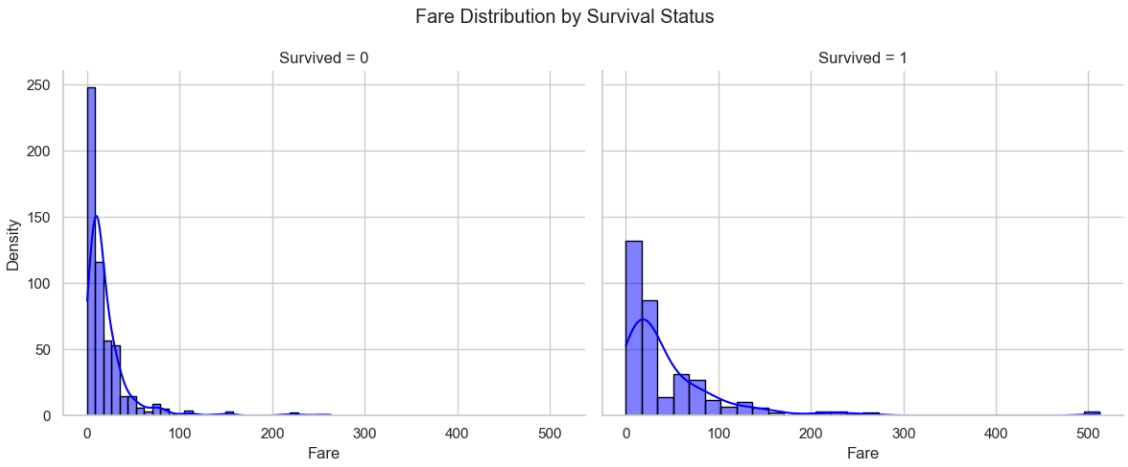

Fig 8 Fare distribution by Survival status

The data presented in the figure suggested that the survival rates for individuals who purchased expensive tickets remained relatively stable and showed no significant variation. In contrast, the survival rates for those who bought cheaper tickets or were in the lower class was markedly lower. This significant gap highlighted a clear difference in survival chances based on the ticket price. The lower class passengers were at much higher risk, indicating the critical role that socioeconomic status played in their chances of survival. This pattern clearly demonstrated how financial resources and class influenced who survived.

4.2 Survival Rates Distribution by Categorical Factors

4.2.1 Sex and Survival Rats

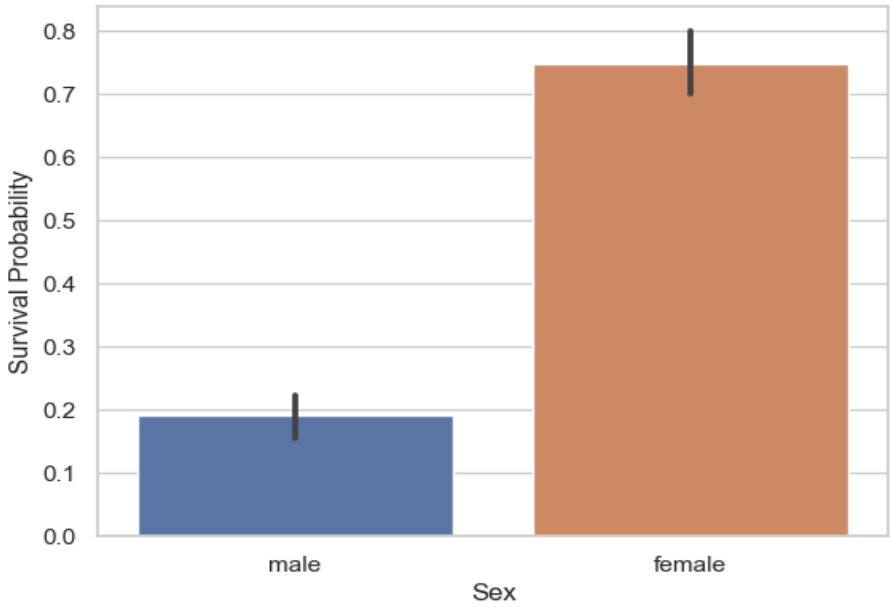

Fig 9 Impact of Sex on Survival rates

It was clear from the figure that males had a much lower chance of survival compared to females on the Titanic. This suggested that sex played a crucial role in determining who survived. For those familiar with the 1997 Titanic movie, the phrase "Women and children first" during the evacuation was memorable and highlighted the gender-based survival priority at the time, influenced by societal norms and evacuation protocols, resulting in higher survival rates for females and underscored the importance of sex as a key factor in survival analysis.

4.2.2 Pclass and Survival Rates

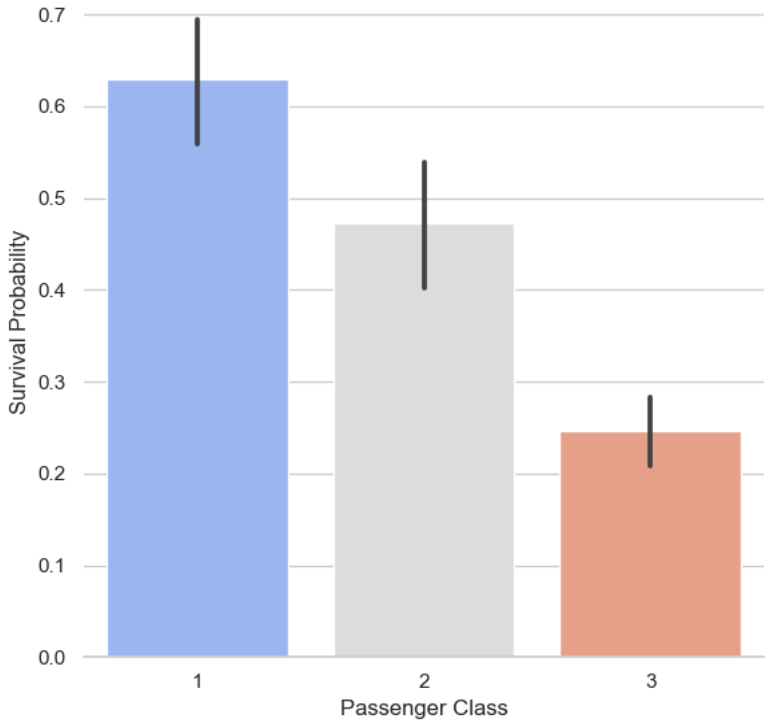

Fig 10 Impact of Passenger Class (Pclass) on Survival rates

The survival rates among Titanic passengers also varied significantly across the three different classes. First-class passengers had the highest chance of survival, followed by second-class passengers, while third-class passengers had the lowest survival rates. Several factors contributed to this disparity, such as the location of cabins, access to lifeboats, and socioeconomic status. First-class passengers, often located closer to the lifeboats and had better access to evacuation routes, leading to quicker and more efficient evacuation. In contrast, third-class passengers, who were on the lower decks, encountered greater obstacles in reaching safety, such as longer distances and possible language barriers. This significant difference in survival rates highlighted the impact of socioeconomic status and class on survival outcomes during the Titanic disaster.

4.2.3 Embarked and Survival Rates

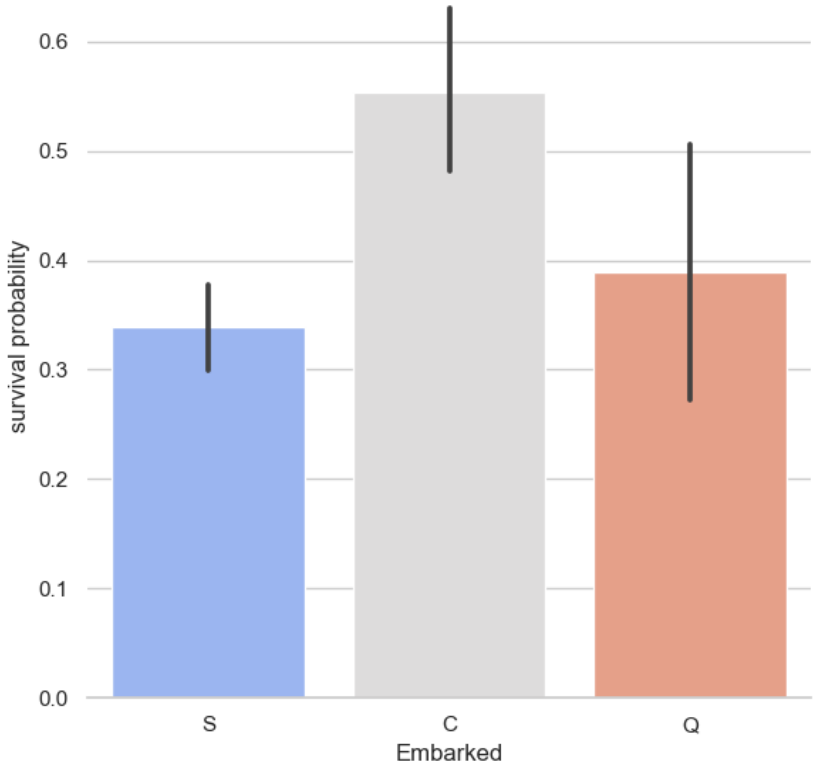

Fig 11 Impact of Embarked on Survival rates

The bar chart showed that passengers who boarded at Cherbourg had noticeably higher survival rates compared to those who embarked at Southampton and Queenstown. At first glance, this difference might seem like a coincidence. However, it deserved further investigation to identify any underlying factors that may be contributing to this significant gap in survival rates. Understanding why Cherbourg passengers did better could provide valuable insights into differences in such as socio-economic status, cabin locations, or other variables that might had affected their chances of survival. Therefore, it was crucial to investigate this trend further to ensure our analysis accurately reflectd the factors influencing survival rates.

4.3 EDA Conclusion

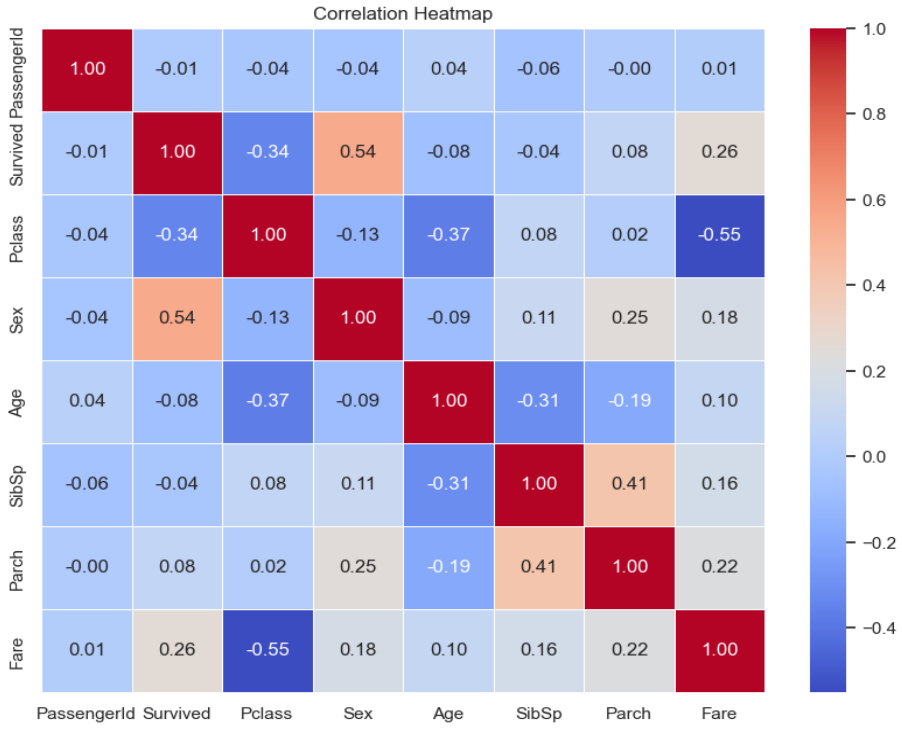

Fig 12 Correlation heatmap among the attributable variables

In conclusion, our analysis indicated that Sex, Pclass, and Fare were the three most critical factors influencing survival rates. These features emerged as the most significant predictors, displaying a strong correlation with the survival rates. While other features such as Age, SibSp, and Parch also played important roles in determining survival rates, their impact was relatively less substantial compared to the key factors identified. Specifically, Sex was a critical factor, with women having higher survival rates than men. Similarly, Pclass revealed a stark contrast in survival chances, with those in higher classes faring better than those in lower classes. Lastly, Fare was also linked to survival, possibly reflecting underlying socio-economic conditions. Therefore, even though acknowledging the significance of various factors, we concluded that Sex, Pclass, and Fare were the main determinants of survival rates.

CHAPTER 5

Model

4.1 Model Evaluation Metrics and Validation

In our model evaluation process, we prioritized accuracy as the main metric for assessing performance. Accuracy measured the proportion of correct predictions out of the total number of observations. In our context, the accuracy of predictions was crucial because it directly affected human lives. In real-world applications such as healthcare, disaster warning systems, and other life safety areas, precise predictions could save lives. Even a small improvement of one percent in accuracy could potentially result in more lives saved when dealing with large volumes of observations. Therefore, accuracy was our top priority when designing and training our models. While accuracy was critical, it was not the only evaluation criterion. We also considered other important factors like robustness, interpretability and fairness of the model. However, we currently believed that accuracy was the most urgent and essential metric because it conducted a direct impact on life and death. By continuously improving the accuracy of our models, we could maximize their life-saving potential.

When training machine learning models, it was crucial to partition the dataset correctly and used appropriate validation methods to ensure the model performs well on new data. Therefore, we split the data into three parts: 70% for training, 20% for testing, and 10%

for validation. Overfitting, where a model learned the training data too well and did not generalize to new data, was a major challenge. Cross-validation helped us assess how well a model would perform on new data and helped prevent overfitting, so we used the 10-fold cross-validation method.

In this study, we compared three different machine learning models: XGBoost, Random Forest, and AdaBoost. The figure below showed the performance of these models during cross-validation. As you could see, the average validation scores of XGBoost and Random Forest were relatively close, both around 80%, while the average validation score of AdaBoost was slightly lower, at about 78%.

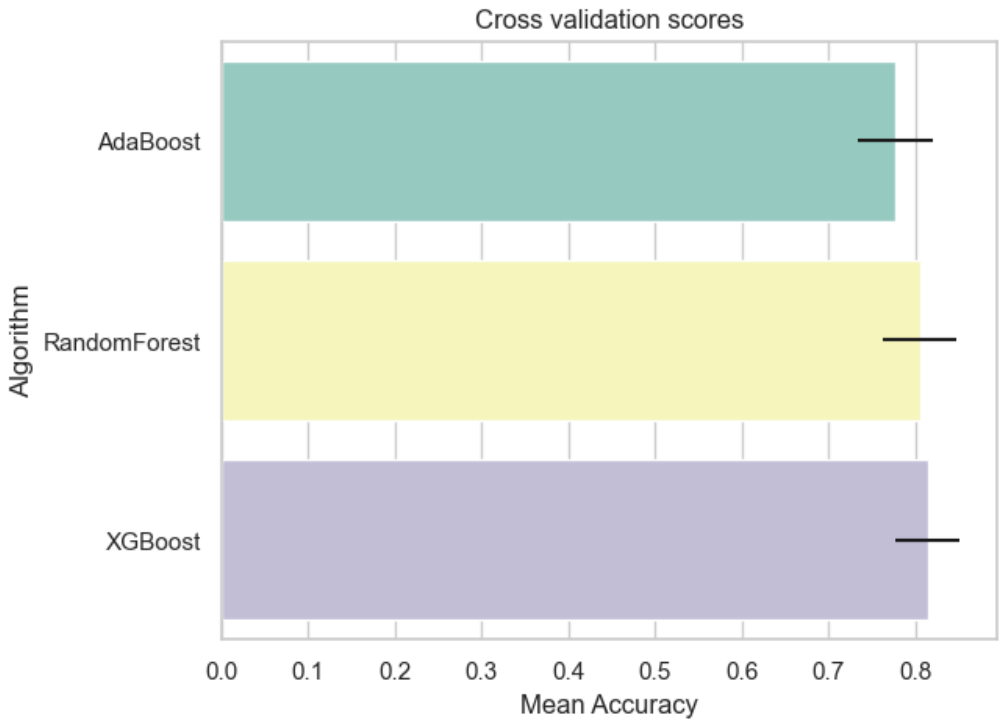

Fig 13 Performance of models during Cross-Validation

4.2 Modeling

To find the optimal parameters for each model, we conducted a full grid search approach. This method allowed us to explore a wide range of parameter combinations to determine the most effective settings for our models. The tables below detailed the specific parameters we tested and the values we found for AdaBoost, Random Forest, and XGBoost.

Parameter	Value
base_estimator__criterion	["gini", "entropy"]
base_estimator__splitter	["best", "random"]
algorithm	["SAMME","SAMME.R"]
n_estimators	[1,2]
learning_rate	[0.0001, 0.001, 0.01, 0.1]

Table 2 Grid search results for parameter optimization of AdaBoost

Parameter	Value
max_depth	none
max_features	[1, 3, 7, 10]
min_samples_split	[2, 5, 10]

min_samples_leaf	[1, 3, 7]
bootstrap	[False]
n_estimators	[50, 100, 300]
criterion	[50, 100,300]

Table 3 Grid search results for parameter optimization of Random Forest

Parameter	Value
max_depth	[3, 5, 7]
learning_rate	[0.01, 0.03, 0.1]
n_estimators	[100,200,300]

Table 4 Grid search results for parameter optimization of XGBoost

4.3 Model Performance

Model	Accuracy
AdaBoost	0.8036516853932584
Random Forest	0.8320480081716036
XGBoost	0.8286006128702759

Table 5 Models performance

From the results above, we observed that the Random Forest model achieved the highest accuracy, coming in at 83.205%. This was followed closely by XGBoost, which attained an accuracy of 82.860%. Among the three models we tested, AdaBoost had the lowest accuracy, registering at 80.37%. Based on these findings, we decided to select the Random Forest model as our final choice. Its superior performance in terms of accuracy made it the most reliable option for our predictive analysis, ensuring that we used the most effective model for our needs. This decision matched our goal of maximizing accuracy and making our final model as robust as possible.

4.4 Feature Importance

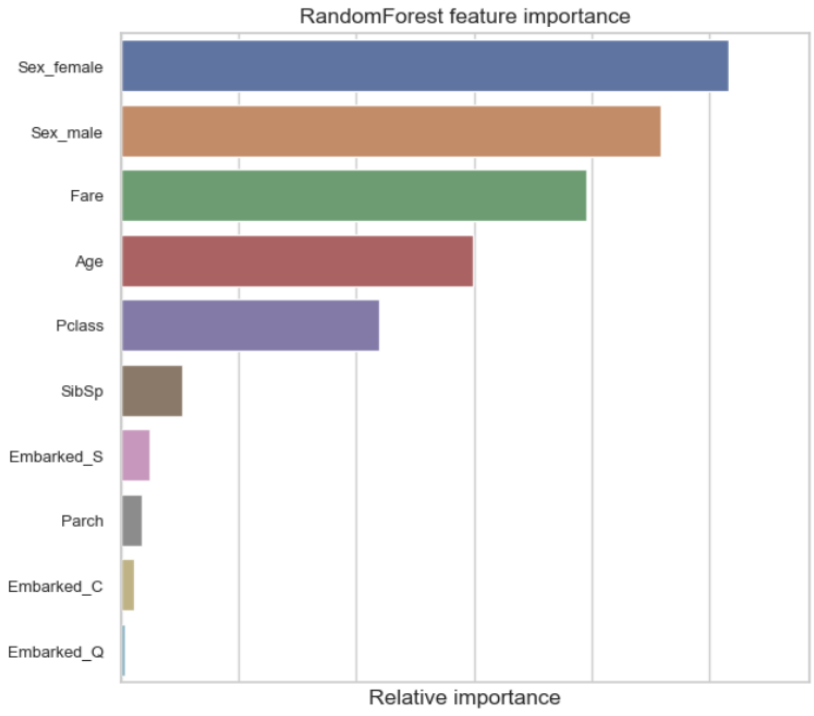

Fig 14 Feature importance analysis of Random Forest for Survival prediction

After selecting the Random Forest model, we analyzed the feature importance to ensure our model not only had high accuracy but also good interpretability. According to the plot,

the most crucial factor for predicting survival was Sex, highlighting the common principle of "ladies first" in emergencies. The next important factor in our model was the Fare, which aligned with our earlier analysis showing that first-class passengers had a higher survival rate, likely due to better facilities and easier access to escape routes. The third one was the Age, which was closely related to gender, as the elderly and children were often typically given priority in emergencies. The fourth key factor was the Pclass, which correlated with the Fare and indicated the survival benefits of higher classes. The other features had minimal impact on our model and could be considered negligible.

CHAPTER 6

Conclusion

In this study, we compared three popular machine learning models: XGBoost, Random Forest, and AdaBoost, and found that the Random Forest model performed the best. Therefore, we recommend employing the Random Forest model to predict survival probabilities during rescue operations after disasters.

In our analysis of feature importance, the variable "Sex" emerged as the most critical factor influencing survival rates. This finding strongly reflected the tradition of "women first" during emergencies, underscoring a deep sense of compassion and responsibility in humanity during crisis. Putting the safety of the most vulnerable first highlighted a noble and admirable aspect of human behavior. Recognizing and valuing this tradition was essential, as it represented a universal moral standard that ensured the welfare and protection of those most in need. We believed that this admirable practice should be cherished and upheld across the globe, serving as a guiding principle in emergency protocols and humanitarian efforts worldwide.

Additionally, the “Pclass” was another important feature. First-class passengers typically had better access to evacuation routes and escape facilities, giving them a significant survival advantage. However, in disasters, the value of human life should not be determined by social status or wealth. Every life was equally important. Therefore, we recommended that shipping companies and ship designers provided more survival

resources to lower-class cabins to ensure that all passengers had an equal chance of survival in emergencies.

In a nutshell, by comparing multiple machine learning models and thoroughly interpreting the results, we developed a high-accuracy survival prediction tool. More importantly, we had highlighted the need to integrate humanitarian concerns and equality into future ship design and disaster response strategies. This approach was not only a technological advancement but also a commitment to respecting life and promoting fairness and justice in our society.

References

1，Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

2，Dhaliwal, Sukhpreet Singh, Abdullah-Al Nahid, and Robert Abbas. "Effective intrusion detection system using XGBoost." *Information* 9.7 (2018): 149.

3，Rigatti, Steven J. "Random forest." Journal of Insurance Medicine 47.1 (2017): 31-39.

4，Speiser, Jaime Lynn, et al. "A comparison of random forest variable selection methods for classification prediction modeling." *Expert systems with applications* 134 (2019): 93-101.

5，Biau, Gérard. "Analysis of a random forests model." *The Journal of Machine Learning Research* 13.1 (2012): 1063-1095.

6，Hastie, Trevor, et al. "Multi-class adaboost." Statistics and its Interface 2.3 (2009): 349-360.

7，Ying, Cao, et al. "Advance and prospects of AdaBoost algorithm." *Acta Automatica Sinica* 39.6 (2013): 745-758.

8，Domingo, C., & Watanabe, O. (2000, June). MadaBoost: A modification of AdaBoost. In *colt* (pp. 180-189).

9，Sneha Bose (2022) “Random Forest Algorithm” Example Website: https://insideaiml.com/blog/Random-Forest-Algorithm-1029